Bayesian Quantile Regression Methods[*]

Tony Lancaster
Department of Economics
Brown University

Sung Jae Jun[†]
Department of Economics and CAPCP[‡]
Pennsylvania State University

First Draft: May 2006
This Version: August 2008

Abstract

This paper is a study of the application of Bayesian Exponentially Tilted Empirical Likelihood to inference about quantile regressions. In the case of simple quantiles we show the exact form for the likelihood implied by this method and compare it with the Bayesian bootstrap and with Jeffreys' method. For regression quantiles we derive the asymptotic form of the posterior density. We also examine MCMC simulations with a proposal density formed from an overdispersed version of the limiting normal density. We show that the algorithm works well even in models with an endogenous regressor when the instruments are not too weak.

Key Words: Bayesian inference: Empirical likelihood: Instrumental variables: Weak identification

[*] We thank two anonymous referees for their helpful comments. Sung Jae Jun also thanks the Human Capical Foundation for their support.

[†] (corresponding author) 619 Kern Bldg, Department of Economics, Pennsylvania State University, University Park PA 16802, sjun@psu.edu, Tel: 814-865-6149, Fax: 814-863-4775

[‡] The Center for the Study of Auctions, Procurements and Competition Policy

1 Introduction

Recent work by Schennach (2005) has opened the way to a new Bayesian treatment of quantile regression. In this paper we shall explain how this method may be applied to quantile regression both when regressor variables are exogenous and when they are endogenous but instrumental variables are available. We give an explicit form for the posterior density of quantiles and a comparison with the method of Jeffreys (1961). We give several examples using both real and artifical data and explore the consquences of having an instrument weakly correlated with the endogenous regressor.

In the remainder of this section we briefly describe previous proposals for Bayesian inference about quantiles and quantile regressions . Section 2 describes Schennach's method and its application to quantiles. In section 3 we compare her method with the Bayesian bootstrap posterior. In section 4 we give the application to quantile regression and in section 5 we give the application of the method to estimation of structural quantile models with endogenous regressors. In an appendix we derive the limiting (normal) form of the posterior density.

The earliest Bayesian method for quantiles that we know of is Jeffreys' substituttion posterior for the median (see Jeffreys (1961), Monahan and Boos (1992), and Lavine (1995)). If n_1 is the number of observations less than or equal to θ and n_0 the number great than θ then Jeffreys pointed out that if θ is the median the probability that n_1 observations are less than it and $n_0 = n - n_1$ are greater is ${}_nC_{n_1}2^{-n}$ and so proposed

$$p(\theta|\text{data}) \propto \frac{1}{n_1!n_0!} \tag{1}$$

as a posterior density for the median, assuming a uniform prior.[1] Note here that this proposal is not based on the distribution of the data but on the distribution of a function of the data *and* the parameter. Therefore, it is not a valid posterior in the sense that it does not follow from Bayes' rule (Monahan and Boos (1992)).[2] Nonetheless, this proposal has been

[1] We owe this reference to Jeffreys' proposal to Roger Koenker.

[2] Monahan and Boos (1992) pointed out that the normalising constant is not "the marginal of any recognizable quantity."

studied by e.g. Lavine (1995) and Dunson and Taylor (2005) as an approximation to a valid posterior. Following Lavine (1995) and Dunson and Taylor (2005), we call this proposal Jeffreys' substitution posterior or simply Jeffreys' posterior. This is a step function. We shall show in section 3 that Schennach's method for the median is numerically very close to Jeffreys' even for quite small n.

Jeffreys' argument would naturally lead to

$$p(\theta_\tau|\text{data}) \propto \frac{\tau^{n_1}(1-\tau)^{n_0}}{n_1!n_0!} \propto \frac{\phi^{n_1}}{n_1!n_0!}, \quad \phi = \frac{\tau}{1-\tau} \tag{2}$$

as an approximate posterior density for the τ'th quantile θ_τ with n_1 as the number of observations less than or equal to θ_τ and n_0 the number greater than it. Lavine (1995) extended Jeffreys' approach to a vector of quantiles, replacing the binomial expression underlying (2) with a multinomial. Dunson and Taylor (2005) in turn extended Lavine's work to handle a vector of quantile regression functions and proposed a Markov Chain Monte Carlo (MCMC) algorithm for sampling the posterior. Yu and Moyeed (2001) propose Bayesian inference about quantile regressions using as a likelihood an asymmetric Laplace distribution for the error term u in a linear model of the form

$$p(u) \propto \exp\{-\rho_\tau(u)\}$$

where $\rho_\tau(u)$ is the check function $\rho_\tau(u) = u(\tau - 1(u \leq 0))$ and $u = Y - X'\beta_\tau$. This reduces to $p(u) = \exp\{-|u|\}$ for median regression. They suggest an MCMC algorithm for sampling the posterior. Kottas and Gelfand (2001) propose median regression using nonparametric median zero distributions for the error term in a linear model. Chamberlain and Imbens (2003) note that it is straightforward to compute the Bayesian bootstrap posterior density in quantile regression by repeatedly solving the problem

$$\beta_\tau = \arg\min_t \sum_{i=1}^n \rho_\tau(v_i u_i), \tag{3}$$

where the $\{v_i\}$ are iid unit exponential variates and $u_i = Y_i - X_i't$. (This problem may be solved, for example, using Koenker's quantile regression R function with unit exponential weights as rq(y˜x, tau, weights=rexp(n).)

2 Bayesian Exponentially Tilted Empirical Likelihood

Schennach's (2005) method, called Bayesian exponentially tilted empirical likelihood (Betel), provides a likelihood for randomly sampled data $\mathbb{Y} = (Y_1, \cdots, Y_n)'$ subject only to a set of m moment conditions of the form $E\big(g(Y_i, \theta)\big) = 0$ where θ is a k dimensional parameter of interest and k may be smaller, equal to or larger than m and $g(., \theta)$ is a vector of known functions. The method may be thought of as construction of a likelihood supported on the n data points that is minimally informative, in the sense of maximum entropy, subject to the moment conditions. Specifically the probabilities $\{p_i\}$ attached to the n data points are chosen to solve

$$\max_{p_1, \cdots, p_n} \sum_{i=1}^{n} -p_i \log p_i \quad s.t. \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(Y_i, \theta) = 0. \tag{4}$$

The solutions of this problem, well known in the maximum entropy literature, e.g. Jaynes (2003, p. 357), take the form

$$p_i(\theta) = \frac{\exp\{\lambda(\theta)' g(Y_i, \theta)\}}{\sum_{j=1}^{n} \exp\{\lambda(\theta)' g(Y_i, \theta)\}} \tag{5}$$

where m vector λ, dependent on θ, satisfies

$$\lambda(\theta) = \arg\min_{\eta} n^{-1} \sum_{i=1}^{n} \exp\{\eta' g(Y_i, \theta)\}. \tag{6}$$

The $\{\lambda_j\}$ are the Lagrange multipliers corresponding to the m constraints in the problem (4). For every θ solving (6) is a convex minimization problem and computationally straightforward.

The resulting likelihood for i.i.d data is $\prod_{i=1}^{n} p_i(\theta)$ and this may be combined with a prior density on θ to yield the posterior density

$$p(\theta|\mathbb{Y}) = p(\theta) \prod_{i=1}^{n} p_i(\theta) \tag{7}$$

on support such that 0 is in the interior of the convex hull of the union of the $\{g(Y_i, \theta)\}$.[3]

3 Posterior Inference about Quantiles

Consider first Bayesian inference about quantiles for which an explicit representation of the Betel posterior is available. The τ'th quantile θ_τ satisfies the moment restriction $E(g_i(\theta_\tau)) = 0$ where

$$g_i(\theta_\tau) = 1\{Y_i \leq \theta_\tau\} - \tau \tag{8}$$

and where $1\{A\}$ is the indicator function of the event A. Given a random sample $\mathbb{Y} = (Y_1, \cdots, Y_n)'$ from a distribution $F(\cdot)$ and for any θ_τ in the support of the Betel likelihood which in this case is $[\min Y_i, \max Y_i)$, the Lagrange multiplier λ solving the problem (6) satisfies the equation

$$\sum_{i=1}^{n} g_i(\theta_\tau) e^{\lambda g_i(\theta_\tau)} = 0$$

with solution

$$e^{\lambda(\theta_\tau)} = \frac{\tau n_0}{(1-\tau) n_1},$$

where $n_1 = \sharp(Y_i \leq \theta_\tau)$ and $n_0 = n - n_1$. Substituting this solution into the expression for the posterior density (5) and assuming a uniform prior gives

$$p(\theta_\tau|\mathbb{Y}) \propto \frac{\phi^{n_1}}{n_1^{n_1} n_0^{n_0}}, \quad \phi = \frac{\tau}{1-\tau}. \tag{9}$$

Like Jeffreys' substitution posterior, the Betel posterior (9) is a piecewise constant density. It is supported on $[\min Y_i, \max Y_i)$. For the median (9) gives $p(\theta_{0.5}|\mathbb{Y}) \propto 1/n_1^{n_1} n_0^{n_0}$ which may

[3] Chernozhukov and Hong (2003) proposed a quasi–posterior using a classical GMM or GEL objective function. However, we emphasize that equation (7) is a posterior derived from Bayes' rule followed by taking the limit that the number of nuisance parameters goes to infinity. In particular, the normalising constant is a limit of marginals with nuisance parameters. For more discussion, see Schennach (2005).

be compared with Jeffreys' $1/n_1!n_0!$.

Figure 1 shows the posterior density of the median from a random sample of size $n = 50$ using a uniform prior. The black step function is (9) and the red is Jeffreys' posterior. It can be seen that the two distributions are very similar. Both Jeffreys' and Betel posterior distributions of quantiles are always step functions with steps at the distinct observations. Highest posterior density intervals with (possibly approximate) 95% content are straightforward to construct.

Although Jeffreys' and the Betel posteriors look quite similar, a formal comparison of the two is subtle and needs more care; the ratio of the two posteriors does not pointwisely coverge to 1 so that they are generally asymptotically different. However, the two posteriors look similar and provide similar inference, because both of them are consistent and they behave similarly in the neighborhood of the true quantile.

To be more precise, let $p_J(\theta_\tau|\mathbb{Y})$ and $p_B(\theta_\tau|\mathbb{Y})$ be Jeffreys' proposal and the Betel posterior at θ_τ, respectively, where θ_τ is such that $0 < F(\theta_\tau) < 1$. Then, Stirling's approximation applied to $n_1!$ and $n_0!$ gives

$$\frac{p_B(\theta_\tau|\mathbb{Y})}{p_J(\theta_\tau|\mathbb{Y})} \propto \frac{n_1!n_0!}{n_1^{n_1}n_0^{n_0}} \propto \sqrt{n_1 n_0}\left(1 + O(\frac{1}{n_1})\right)\left(1 + O(\frac{1}{n_0})\right), \tag{10}$$

where θ_τ is guaranteed to be on the support of the two posteriors if n is large enough. Apply the central limit theorem and the Delta method to n_1/n and n_0/n in (10), and we have an asymptotic expression

$$p_B(\theta_\tau|\mathbb{Y}) = C_n p_J(\theta_\tau|\mathbb{Y})\left(\sqrt{F(\theta_\tau)(1 - F(\theta_\tau))} + O_p(\frac{1}{\sqrt{n}})\right) \tag{11}$$

for each θ_τ such that $0 < F(\theta_\tau) < 1$ and for some C_n that does not depend on θ_τ. Therefore,

$$\frac{p_B(\theta_\tau|\mathbb{Y})}{C_n p_J(\theta_\tau|\mathbb{Y})} \xrightarrow{p} \sqrt{F(\theta_\tau)(1 - F(\theta_\tau))} \neq 1.$$

Note that this is not a rescaling issue, because the limit in the right–hand side depends on the parameter value θ_τ.

Then, why does Jeffreys' proposal look so similar to the Betel? In order to answer this quation, let θ_τ^0 be the true quantile so that $F(\theta_\tau^0) = \tau$. It then follows from equation (11) that

$$\frac{p_J(\theta_\tau|\mathbb{Y})}{p_J(\theta_\tau^0|\mathbb{Y})} = \frac{p_B(\theta_\tau|\mathbb{Y})}{p_B(\theta_\tau^0|\mathbb{Y})} \cdot \frac{\sqrt{\tau(1-\tau)} + O_p(\frac{1}{\sqrt{n}})}{\sqrt{F(\theta_\tau)(1-F(\theta_\tau))} + O_p(\frac{1}{\sqrt{n}})}. \tag{12}$$

Since the Betel posterior is consistent, as is proved in the appendix, equation (12) shows that Jeffreys' proposal is also consistent in the sense that $p_J(\theta_\tau|\mathbb{Y})/p_J(\theta_\tau^0|\mathbb{Y}) \xrightarrow{p} 0$ when $\theta_\tau \neq \theta_\tau^0$.

Moreover, equation (12) shows that Jeffreys' proposal provides almost the same inference as the Betel posterior when θ_τ is not too far from θ_τ^0. That is, assuming that $F(\cdot)$ is continuous at θ_τ^0, we have

$$\frac{p_B(\theta_\tau^0|\mathbb{Y})}{p_B(\theta_\tau|\mathbb{Y})} \cdot \frac{p_J(\theta_\tau|\mathbb{Y})}{p_J(\theta_\tau^0|\mathbb{Y})} = \frac{\sqrt{\tau(1-\tau)} + O_p(\frac{1}{\sqrt{n}})}{\sqrt{F(\theta_\tau)(1-F(\theta_\tau))} + O_p(\frac{1}{\sqrt{n}})} \approx 1, \tag{13}$$

which shows that Jeffreys' substitution posterior distinguishes θ_τ^0 from θ_τ almost as well as the Betel posterior when θ_τ is close to θ_τ^0. We summarize this result in the following proposition.

Proposition 1 *Let $F(\cdot)$ be the true distribution that is continuous at the true quantile θ_τ^0. For any $\theta_\tau \neq \theta_\tau^0$ such that $0 < F(\theta_\tau) < 1$, we have*

$$\frac{p_J(\theta_\tau|\mathbb{Y})}{p_J(\theta_\tau^0|\mathbb{Y})} \xrightarrow{p} 0.$$

Moreover, for any $\epsilon > 0$, there exists $\delta > 0$ such that $||\theta_\tau^0 - \theta_\tau|| < \delta$ implies

$$\Big| \log\Big(\frac{p_B(\theta_\tau^0|\mathbb{Y})}{p_B(\theta_\tau|\mathbb{Y})} \cdot \frac{p_J(\theta_\tau|\mathbb{Y})}{p_J(\theta_\tau^0|\mathbb{Y})}\Big)\Big| \leq \epsilon + o_p(1).$$

Proof: It follows from equation (12).

Proposition 1 is comparable to Lavine (1995), where he showed that Jeffreys' proposal provides conservative inference in large samples relative to the true (unknown) posterior. In particular, he showed that when $\theta_\tau^0 \neq \theta_\tau$,

$$\liminf_{n\to\infty} \log\Big(\frac{p_J(\theta_\tau|\mathbb{Y})}{p_J(\theta_\tau^0|\mathbb{Y})} \cdot \frac{\ell(F)}{\ell(\tilde{F})}\Big) \geq 0 \tag{14}$$

with probability one under F, where F is the true distribution with its τ^{th} quantile θ_τ^0, $\tilde{F}$ is an alternative distribution with its τ^{th} quantile θ_τ, and $\ell(F)$ denotes the likelihood based on F. Lavine (1995) interpreted inequality (14) as the one showing that Jeffreys' proposal distinguishes between θ_τ^0 and θ_τ less well than the true (unknown) likelihood and hence it leads to conservative inference. Proposition 1 shows that Jeffreys' proposal distinguishes θ_τ^0 from θ_τ as well as the Betel posterior which is essentially semiparametric. Since Lavine compared Jeffreys' proposal to inference based on the true unknown likelihood, he was comparing semiparametric inference to parametric one. Both Jeffreys' and Betel are semiparametric methods and we have found that they are as good as each other.

4 Comparison with the Bayesian Bootstrap

The Bayesian bootstrap of Rubin (1981) takes the data to be iid multinomial with probabilities $\{p_i\}$. An improper Dirichlet prior on these probabilities leads to a Dirichlet posterior that assigns positive probability only to the distinct sample observations and in this respect is similar to Betel. Parameters such as quantiles that can be expressed as functionals of the data distribution have posterior distributions that can be calculated by repeatedly simulating from the Dirichlet posterior distribution of the $\{p_i\}$ and calculating the parameter of interest. As indicated in section 1 this amounts to repeatedly solving the problem (3).

To compare the Bayesian bootstrap (BB) and Betel (uniform prior) posteriors we consider inference about the median – $\tau = 0.5$ – using a sample $n = 500$ standard normal variates. To get the BB posterior we solved the problem (3) 10,000 times using the method described in section 1 and drew the histogram of the realizations. This is shown in figure 2. Note the sparsity of the Bayesian bootstrap distribution which reflects the fact that there were only 162 distinct realizations among the 10,000 draws even though the sample size was 500. This arises because the criterion function in (3) is a piecewise linear function with knots at the data points so solutions of the problem will always lie at one of the data points. Hence there can be at most n points of support for the Bayesian bootstrap distribution and with $n = 500$ most of these will have probability so low that they will not occur in a sample of 10,000 realizations. By contrast, the Betel distribution, shown in red, provides positive probability density over the relevant interval.

5 Quantile Regressions

Consider a simple τ'th quantile regression with a single regressor

$$\Pr(Y \leq \alpha(\tau) + \beta(\tau)X|X) = \tau,$$

which implies unconditional moment condtions

$$E\big(1\{Y \leq \alpha(\tau) + \beta(\tau)X\} - \tau\big) = E\big(X(1\{Y \leq \alpha(\tau) + \beta(\tau)X\} - \tau)\big) = 0.$$

If we now define

$$g_{1i} = 1\{Y_i \leq \alpha(\tau) + \beta(\tau)X_i\} - \tau \qquad \text{and} \qquad g_{2i} = X_i(1\{Y_i \leq \alpha(\tau) + \beta(\tau)X_i\} - \tau),$$

we may compute the Lagrange multipliers λ_1 and λ_2 by

$$\lambda = \arg\min_{\eta} \sum_{i=1}^{n} \exp\{\eta_1 g_{1i} + \eta_2 g_{2i}\}$$

and then calculate the posterior density according to (5).

Example 1: Demand for Fish. Figure 3 plots the joint posterior density of $\alpha(0.5)$ and $\beta(0.5)$ from a sample of $n = 111$ observations and under a uniform prior. The data are Graddy's (1995) fish market observations with Y as log quantity traded and X as log price. To construct figure 3 the density was evaluated on a 100×100 grid of α, β values. As the figure shows the density consists of adjoining flat surfaces. The marginal densities of α and β may be calculated by summing this grid across rows or across columns. Figure 4 shows the marginal density of the price elasticity of demand at the 0.5 quantile found in this way. The quantile regression estimate of $\beta(0.5)$ was -0.41 which can be seen to be close to the marginal posterior mode. (The quantile regression was computed using rq(Q˜P) in R, where Q and P are the logarithms of quantity and price).

6 Quantiles with Endogenous Covariates

Quantile regression applied to the observations on price and quantity neglects the simultaneity of these variables when the market is in equilibrium. This can be surmounted by use of instrumental variables.

Following Chernozhukov and Hansen (2006) consider the model

$$Y = D'\alpha(U) + X'\beta(U), \quad U|X, Z \sim \text{Uniform}(0, 1)$$

in which D is statistically dependent on U, $D'\alpha(\tau) + X'\beta(\tau)$ is strictly increasing in τ, and Z is a set of instrumental variables that are independent of U but statistically dependent on D. Then $D'\alpha(\tau) + X'\beta(\tau)$ is the τ'th quantile of Y conditional on X, Z. That is,

$$\Pr(Y \leq D'\alpha(\tau) + X'\beta(\tau)|X, Z) = \tau \tag{15}$$

The expression

$$D'\alpha(\tau) + X'\beta(\tau)$$

is what Chernozhukov and Hansen refer to as a structural quantile function. The fact (15) then leads to unconditional moments of which the simplest are of the form

$$E\big(X_i(1\{y_i \leq D_i'\alpha(\tau) + X_i'\beta(\tau)\} - \tau)\big) = 0 \quad \text{and} \quad E\big(Z_i(1\{y_i \leq D_i'\alpha(\tau) + X_i'\beta(\tau)\} - \tau)\big) = 0.$$

We may then apply the Betel method using these moment functions. We illustrate first using simulated data designed to capture the behavior of the posterior both when the instruments, Z, are strongly correlated with the included endogenous variable D and when they are weakly correlated with it, a situation known to lead to difficulties in linear models.

Example 2: Simulated Data: In the following example the data were generated with $X = 1$ and eight instruments which are the columns of Z drawn from $N(0,\ I_8)$. The Y data were generated by

$$Y_i = D_i \alpha(U_i) + \beta(U_i) \quad \text{and} \quad D_i = \gamma_0 + Z_i \gamma_1 + V_i,$$

$$\begin{bmatrix} \beta(U_i) \\ V_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right).$$

τ was set equal 0.25 and $n = 500$ was used. We specified $\alpha(s) = 1$ for all $s \in (0, 1)$. For the first experiment the eight elements of γ_1 were set equal to 1 and in the second they were set equal to 0.1 The latter choice was intended to represent weak instruments. Figures 5 and 6 show the joint posterior densities of $\alpha(0.25)$ and $\beta(0.25)$, the slope and intercept of the structural quantile function at the 0.25 quantile. Figure 4, with $\gamma_1 = 1$ shows a well behaved joint posterior density centered round the truth. Figure 5, with weaker instruments, shows a thicker tailed distribution with apparent multiple modes. Further experiments not shown here show that as the coefficients on the instruments approach zero the joint density shows many modes and thick tails.

Example 3: Demand for Fish Revisited: In this example we again use Graddy's data, also used by Chernozhukov and Hansen. Specifically, we use 111 observations on quantities of fish traded and their price. We also use observations on two weather variables which might be supposed to affect the supply of fish but not the demand. These are called "stormy" and "mixed."

The τ'th structural quantile function is

$$Q = \alpha(\tau)P + \beta(\tau)$$

and this is estimated using three moment equations corresponding to stormy, mixed and the unit variable.[4] Figure 7 shows the joint posterior density of $\alpha(0.5)$ and $\beta(0.5)$. The density was evaluated on a 40×40 grid. It can be seen that there is only limited evidence of weak instruments in the suggestion of a secondary mode. The marginal densities can be found by summing over rows or columns and renormalising. The marginal density of the slope – elasticity of demand – is shown in figure 7. An approximate 95% highest posterior density interval runs from 0.1 to -2.5 and is marked in red. Chernozhukov and Hansen report a point estimate of -0.9 (marked in blue) using these same instruments with a 95% confidence interval running from 0 to -1.8. Note the minor mode in the marginal posterior density.

7 Markov Chain Monte Carlo with Endogenous Regressors

In this section we describe our experiments with Markov Chain Monte Carlo (MCMC) sampling of the joint posterior in an instrumental quantile model. The computational experiments show that a Metropolis-Hasting (MH) algorithm using as a proposal density an overdispersed version of the limiting posterior density works well except when identification is weak. In this case the Betel posterior may have multiple local modes and/or pathologically heavy tails and it seems difficult to choose an appropriate proposal density. When the target distribution has multiple modes and/or heavy tails, the performance of an MCMC is extremely sensitive to the choice of a proposal distribution and the chain may capture only a local feature of the target distribution, or the convergence can be too slow to be practical.

Schennach (2005) showed that the posterior of $\sqrt{n}(\theta - \theta_0)$ converges to a multivariate normal distribution with mean zero and variance matrix equal to $(G'\Omega^{-1}G)^{-1}$, where

[4] Graddy's dataset also contains variables for on–shore weather conditions, and including them as exogenous regressors will make the off-shore weather variables more plausible instruments. Note please that the main purpose here is in an illustration of the method.

$G = E(\partial g(y, \theta_0)/\partial\theta')$ and $\Omega = E(g(y, \theta_0)g(y, \theta_0)')$. Her derivation of the asymptotic form of the posterior assumed differentiablity of the moment function and so does not cover quantile problems. However, as we show in the appendix, a similar result obtains for quantile problems. Specifically, the Betel posterior can be approximated in large samples by a multivariate normal density with mean zero and variance equal to $(G'\Omega^{-1}G)^{-1}$, where $G = \partial E(g(y, \theta_0))/\partial\theta'$ and $\Omega = E(g(y, \theta_0)g(y, \theta_0)')$. (Note that we need the differentiability of the expectation of $g(y, \theta)$, not the differentiability of $g(y, \theta)$ itself.) Although computing a good estimate of G could be difficult in practice, several methods are available, and a kernel method is one of them. We give the method that we used in the algorithm below.

In the following experiments, we consider an instrumental quantile model with various qualities of identification. Specifically, we use the same setup as example 2 with various values of γ_1. We set all eight elements of γ_1 to be 1, 0.1, 0.08, or 0.05 to consider situations of weak instruments. We consider $n = 500$ and $\tau = 0.5$. Although one would want to use MCMC when the parameter of interest is relatively high dimensional, for these experiments we work with a two dimensional parameter, $[\alpha(0.5),\ \beta(0.5)]'$ so that we may compare the exact posterior with its Monte Carlo estimate.

The simulation design in example 2 implies the following moment condition:

$$m(a,b) = \begin{bmatrix} m_1(a,b) \\ m_2(a,b) \end{bmatrix} = \begin{bmatrix} E(1\{Y \le Da + b\} - \tau) \\ E(Z(1\{Y \le Da + b\} - \tau)) \end{bmatrix} = 0$$

when $a = \alpha(\tau)$ and $b = \beta(\tau)$. To obtain samples from the posterior, we used the following MH algorithm.

Algorithm

1. Calculate $\widehat{\alpha}(\tau)$ and $\widehat{\beta}(\tau)$ as follows.

   (a) For each $a \in R$, calculate $\widehat{b}(a)$ by solving $\min_b \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - D_i a - b)$.

   (b) Find $\widehat{\alpha}(\tau)$ by solving $\min_a \parallel \frac{1}{n}\sum_{i=1}^{n} Z_i(1\{Y_i \le \widehat{b}(a) + D_i a\} - \tau) \parallel^2$, which can be solved by a one-dimensional grid search.

(c) Find $\widehat{\beta}(\tau) = \widehat{b}(\widehat{\alpha}(\tau))$, which be read from the data in steps 1 (a) and (b).

2. Calculate $\widehat{V} = (\widehat{G}'\widehat{\Omega}^{-1}\widehat{G})^{-1}$:

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g_i(\widehat{\theta})g_i(\widehat{\theta})' \qquad \text{and} \qquad \widehat{G} = \frac{1}{nh_n}\sum_{i=1}^{n}\begin{bmatrix} k_{\tau,i} & D_i k_{\tau,i} \\ Z_i k_{\tau,i} & Z_i D_i k_{\tau,i} \end{bmatrix},$$

   where $\widehat{\theta} \equiv \left[\widehat{\alpha}(\tau), \widehat{\beta}_1(\tau)\right]'$, $g_i(\theta) \equiv g_i(a,b) \equiv [1, Z_i']' \left(1\{Y_i \le D_i a + b\} - \tau\right)$, and $k_{\tau,i} \equiv k(\frac{Y_i - \widehat{\beta}(\tau) - D_i\widehat{\alpha}(\tau)}{h_n})$. Here, $k(\cdot)$ is a pdf type kernel, and h_n is a bandwidth choice. We used the density of the standard normal as a kernel, and we set h_n to be proportional to $n^{-\frac{1}{5}}$. (Specifically, we set $h_n = cn^{-\frac{1}{5}}$, where c is the interquantile range of $\widehat{\epsilon}_i = Y_i - \widehat{\beta}(\tau) - D_i\widehat{\alpha}(\tau)$.)

3. Choose an initial value θ_j such that $p(\theta_j) > 0$, where $p(\cdot)$ is the Betel posterior. For this purpose, we drew θ_j from $N(\widehat{\theta}, \frac{\widehat{V}}{n})$ and checked if $p(\theta_j) > 0$.

4. Draw θ^* from $N(\theta_j, \frac{\widehat{V}}{n}\kappa^2)$, where κ is a rescaling multiple chosen by the researcher.

5. Calculate $r = \min(1, \frac{p(\theta^*)}{p(\theta_j)})$, where $p(\theta)$ is the Betel posterior.

6. Set $\theta_{j+1} = \begin{cases} \theta^* \text{ with probability } r \\ \theta_j \text{ otherwise} \end{cases}$.

7. Increment j and go to (4).

8. (Optional Adjustment) After some simulations, adjust $\frac{\widehat{V}}{n}$ to the variance matrix estimated from the simulations. Then go back to (4).

This algorithm needs some comments. Step 1 is for calculating an estimate of θ_0. Since it can be difficult to directly calculate the mode of the Betel posterior, we take an indirect way of computation. Since all we need is a reasonably good estimate of θ_0, we simply minimize the quadratic form of the sample analogue of the concentrated moment condition, $m_c(a) \equiv m_2(a, b(a))$, where $b(a)$ solves $m_1(a,b) = 0$ for each a. By using the concentrated moment condition, the computation can be decomposed into a linear programming part (step a) and a one-dimensional grid search part (step b). Even when there are several exogenous variables, the computation of this step is sufficiently easy. Step 2 calculates the variance of the asymptotic Betel posterior. Note that $\widehat{G}$ is an estimate of $G = \partial E(g(x, \theta_0))/\partial\theta' = E([1,\ Z_i']' \, [f_\epsilon(0 \mid D_i, Z_i),\ D_i f_\epsilon(0 \mid D_i, Z_i)])$, where $f_\epsilon(\cdot \mid D, Z)$ is the conditional density of $\epsilon = Y - \beta(\tau) - D_i\alpha(\tau)$ given D and Z. Steps $4-7$ describe an MH sampler. The rescaling parameter κ is chosen by the researcher by checking the acceptance rate of the chain. Although this step may need some trial–and–error in practice, we simply used $\kappa = 2.4/\sqrt{2}$, as recommended in Gelman *et al.* (1995, p. 334).

Following the setup of example 2, we simulated artificial data by setting all eight elements of γ_1 to $1, 0.1, 0.08$, and 0.05. With sample size $n = 500$, F statistics for the significance of the instruments were $535.5, 9.0, 5.5$, and 2.9, respectively. The p values of these statistics are all less than 1%. The values of γ_1 are intended to consider situations of strong and weak instruments. For each simulated dataset, we ran 5 independent chains of $10,000$ iterations, and collected the last $1,000$ realizations from each chain.(So the size of an MCMC sample is $5,000(= 5\times1,000)$.) Figure 9 illustrates the posteriors and MCMC samples for the parameter $\alpha_1(0.5)$ under various values of γ_1. The true marginal posteriors were calculated by numerical integration and rescaling of the joint posteriors.

As we artificially made instruments weaker and weaker, the Betel posteriors started showing multiple modes and heavy tails. The MCMC samples using our algorithm could capture the posteriors until their tails became too thick. In particular, while panels (a)-(b) in figure 9 show a reasonable performance of the algorithm, panels (c)-(d) suggest that it may take extremely long for the chain to converge. The slow convergence can be explained by the fact that we used a normal distribution as a jumping rule when the target distributions were multimodal and fat-tailed. Although a mixture of several fat-tailed distributions could be considered as a proposal distribution, it may not be so practical because checking the pathological shape of a posterior can be computationally infeasible when the parameter of interest is high-dimensional. Therefore, more attention should be paid to checking diagnostics in practice. Although there is no method to *prove or confirm* the convergence of a chain, there are several diagnostics available.[5] For example, running several chains independently and monitoring the autocorrelation functions (ACF) of them can be quite useful. If some of the ACFs are found to be too sticky, then we might suspect the posterior to have a pathological shape. As an example, we provide two pictures of ACFs from our experiments. Figure 10 shows two ACFs for the draws of $\alpha_1(0.5)$, when γ_1 was set to be 0.08. Each panel illustrates an ACF estimated from the last $1,000$ realizations of one of the five chains we ran. We do not provide the ACFs of all five chains, because they were all similar to one of these two pictures. The fact that type I and type II ACFs show up together after $10,000$ iterations can be noted as an indication of a multimodal and/or fat-tailed posterior.

Improving the performance of the algorithm would be an interesting discussion. It may help to use a mixture of several fat-tailed distributions as a jumping rule. Using a prior with a finite support could also be useful. However, all these possibilities seem to require a certain amount of knowledge on the shape of the target posterior.

[5] We emphasize that the slow convergence of a chain does not invalidate the Betel posterior but that it only makes computation more difficult. Note also that the slow convergence of the MCMC can be readily detected as figure 10 shows.

APPENDIX: THE LIMITING POSTERIOR DENSITY

In this appendix, we consider the asymptotic behavior of the Betel posterior under the assumption of identification. In particular, we will show that our log-likelihood admits a quadratic expansion in a local neighborhood of θ_0. Although Schennach (2005) showed that the Betel posterior has a normal approximation in large samples, her derivation assumed differentiabiity of $g(y_i, \theta)$, which does not cover quantile problems. We will derive the quadratic expansion under alternative assumptions. It is also worthy of attention that the approximating normal distribution coincides with the frequentist sampling distribution of an efficient generalized methods of moments (GMM) estimator. The approximation is local in the sense that we only consider a small neighborhood of θ_0. But, it is not a limitation under the assumption of identification because the posterior becomes concentrated around θ_0 and it suffices to consider only a neighborhood of θ_0. The normal approximation of the posterior suggests a candidate jumping distribution in implementing MCMC in practice.

Assumption A *Suppose that data are iid. Let $\mathcal{Y}$ denote the support of y_i.*

(i) $E(g(y_i, \theta)) = 0$ only when $\theta = \theta_0$, where θ_0 is in the interior of the parameter space $\Theta \equiv \big\{\theta \in \mathbb{R}^k : 0$ is in the interior of the convex hull of $\cup_{i=1}^{n} \{g(\bar{y}_i, \theta)\}$ for some $n \in \mathbb{N}$ and some $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n \in \mathcal{Y}\big\}$.

(ii) $\frac{\partial E(g(y_i,\theta))}{\partial \theta'}\big|_{\theta=\theta_0}$ exists, and it has a full column rank.

(iii) $\lambda^*(\theta) \equiv \arg\min_\lambda E(\exp(\lambda' g(y_i, \theta)))$ is differentiable at θ_0. Note that $\lambda^*(\theta_0) = 0$ in view of part (i). We further assume that $\lambda^*(\theta) = 0$ only when $\theta = \theta_0$.

(iv) $\sup_{\theta\in\Theta} \parallel \lambda^*(\theta) \parallel < \infty$ and $\sup_{y\in Y,\theta\in\Theta} \parallel g(y, \theta) \parallel < \infty$.

(v) $E(g(y_i, \theta)g(y, \theta)')$ is continuous at θ_0 and nonsingular.

(vi) $\mathcal{F} = \{\lambda' g(y_i, \theta) : \lambda \in \Lambda, \theta \in \Theta\}$ is a Glivenko-Cantelli and Donsker class with a square-integrable envelope function.

Parts (i), (ii), and (iii) assume identification. Part (iv) assumes a bounded moment function. In quantile regression, it is satisfied when support of the regressors is bounded. The assumption of a bounded moment function is commonly taken in the robust estimation literature (e.g., Huber (1964)). Part (vi) imposes a restriction on $\mathcal{F}$ but it allows the function $g(y_i, \theta)$ to be non-smooth with respect to θ. For example, letting $y_i = (Y_i, X_i', Z_i')'$, $g(y_i, \theta) = Z_i(1\{Y_i \leq X_i'\theta\} - \tau)$ is well-known to satisfy this assumption (see e.g. van der Vaart (1998)). One implication of parts (iv) and (vi) is that $\mathcal{G} = \{\exp(\lambda' g(y_i, \theta)) : \lambda \in \Lambda, \theta \in \Theta\}$ is also Glivenko-Cantelli and Donsker. Therefore, we in fact have

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n g(y_i, \theta) \xrightarrow{p} E(g(y_i, \theta)) \tag{16}$$

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n g(y_i, \theta) g(y_i, \theta)' \xrightarrow{p} E(g(y_i, \theta) g(y_i, \theta)') \tag{17}$$

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \lambda' g(y_i, \theta) \xrightarrow{p} E(\lambda' g(y_i, \theta)) \tag{18}$$

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \exp(\lambda' g_{(}y_i, \theta)) \xrightarrow{p} E(\exp(\lambda' g_{(}y_i, \theta))), \tag{19}$$

all uniformly in θ and λ. Note that it follows from (19) that

$$\lambda_n(\theta) \equiv \underset{\lambda}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \exp(\lambda' g(y_i, \theta)) \xrightarrow{p} \lambda^*(\theta) \tag{20}$$

uniformly in θ. Combining (18) and (19) with (20), we also know that

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \lambda_n(\theta)' g_i(\theta) \xrightarrow{p} E(\lambda^*(\theta)' g_i(\theta)) \tag{21}$$

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \exp(\lambda_n(\theta)' g_i(\theta)) \xrightarrow{p} E(\exp(\lambda^*(\theta)' g_i(\theta))), \tag{22}$$

all uniformly in θ.

Now, we state a proposition that shows consistency and quadratic approximation of the Betel posterior in large samples whose proof uses the convergence results outlined above.

Proposition A *Suppose assumption A holds. Let* $w_i(\theta) \equiv \frac{\exp\left(\lambda_n(\theta)' g(y_i, \theta)\right)}{\frac{1}{n}\sum_{i=1}^n \exp\left(\lambda_n(\theta)' g(y_i, \theta)\right)}$. *Then, for any* $\theta \neq \theta_0$, *we have*

$$\frac{\prod_{i=1}^n w_i(\theta)}{\prod_{i=1}^n w_i(\theta_0)} \xrightarrow{p} 0$$

as $n \to \infty$. *Moreover, for any* $\epsilon > 0$, *there exists a sufficiently small* $\delta > 0$ *such that*

$\| \theta - \theta_0 \| < \delta$ *implies*

$$| \frac{1}{n}\sum_{i=1}^{n} \log w_i(\theta) + \frac{1}{2}(\theta-\theta_0)'\Omega^{-1}(\theta-\theta_0) | \leq \epsilon + o_p(1)$$

as $n \to \infty$*, where* $\Omega^{-1} = (\frac{\partial E(g_i(\theta))}{\partial \theta'}\big|_{\theta=\theta_0})'E(g_i(\theta_0)g_i(\theta_0)')^{-1}(\frac{\partial E(g_i(\theta))}{\partial \theta'}\big|_{\theta=\theta_0})$ *and* $o_p(1)$ *does not depend on* θ*.*

Proof: We will write $g_i(\theta)$ for $g(y_i, \theta)$ for the sake of simplicity. In view of (21) and (22), we first know that

$$\frac{1}{n}\sum_{i=1}^{n} \log w_i(\theta) - \frac{1}{n}\sum_{i=1}^{n} \log w_i(\theta_0) \xrightarrow{p} m(\theta) \equiv E(\lambda^*(\theta)'g_i(\theta)) - \log(E(\exp(\lambda^*(\theta)'g_i(\theta)))),$$

uniformly in θ. Note here that $m(\theta) < 0$ for every $\theta \neq \theta_0$ due to identification and Jensen's inequality. Hence, for every $\theta \neq \theta_0$, we have

$$\frac{1}{n}\log(\prod_{i=1}^{n} w_i(\theta)/\prod_{i=1}^{n} w_i(\theta_0)) \xrightarrow{p} m(\theta) < 0,$$

which is possible only when $\log(\prod_{i=1}^{n} w_i(\theta)/\prod_{i=1}^{n} w_i(\theta_0)) \xrightarrow{p} -\infty$. Therefore, the first part of the proposition obtains.

The second part is obtained by expanding $\frac{1}{n}\sum_{i=1}^{n}\log w_i(\theta)$. For heuristic arguments, note that when θ is close enough to θ_0, we have the following expansion:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \log w_i(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\lambda_n(\theta)'g_i(\theta) - \log(\frac{1}{n}\sum_{i=1}^{n}\exp(\lambda_n(\theta)'g_i(\theta))) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\lambda_n(\theta)'g_i(\theta) - \log(1 + \frac{1}{n}\sum_{i=1}^{n}\lambda_n(\theta)'g_i(\theta) + \frac{1}{2}\lambda_n(\theta)'(\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)')\lambda_n(\theta)) \\ &\approx -\frac{1}{2}\lambda_n(\theta)'(\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)')\lambda_n(\theta) \approx -\frac{1}{2}\lambda^*(\theta)'E(g_i(\theta)g_i(\theta)')\lambda^*(\theta),\end{aligned}$$

which shows that linearizing $\lambda^*(\theta)$ results in the desired approximation. The first approximation is due to the expansion of the exponential function, and the second one is obtained by expanding the logarithm. We will formalize this heuristic argument in several steps. In the following, C will denote a generic constant.

Step 1 (Approximating the exponential part): We claim that for any $\epsilon > 0$, there exists $\delta > 0$ such that $\| \theta - \theta_0 \| < \delta$ guarantees that

$$| \frac{1}{n} \sum_{i=1}^{n} \Big(\exp\big(\lambda_n(\theta)' g_i(\theta)\big) - 1 - \lambda_n(\theta)' g_i(\theta) - \frac{1}{2} \lambda_n(\theta)' g_i(\theta) g_i(\theta)' \lambda_n(\theta) \Big) | \leq \epsilon + o_p(1)$$

as $n \to \infty$, where $o_p(1)$ does not depend on θ.

To prove this claim, recall that $| \exp(x) - 1 - x - \frac{1}{2}x^2 | \leq C \mid x \mid^3$ when x belongs to a bounded interval. Note that part (iv) in assumption A guarantees that $\lambda_n(\theta)' g_i(\theta)$ belongs to a bounded interval for every $\theta \in \Theta$ for sufficiently large n. Therefore, for sufficiently large n, we have

$$\begin{aligned} &| \frac{1}{n} \sum_{i=1}^{n} \Big(\exp\big(\lambda_n(\theta)' g_i(\theta)\big) - 1 - \lambda_n(\theta)' g_i(\theta) - \frac{1}{2} \lambda_n(\theta)' g_i(\theta) g_i(\theta)' \lambda_n(\theta) \Big) | \\ &\qquad \leq \frac{1}{n} \sum_{i=1}^{n} \big| \exp\big(\lambda_n(\theta)' g_i(\theta)\big) - 1 - \lambda_n(\theta)' g_i(\theta) - \frac{1}{2} \lambda_n(\theta)' g_i(\theta) g_i(\theta)' \lambda_n(\theta) \big| \\ &\qquad\qquad \leq C \frac{1}{n} \sum_{i=1}^{n} \big| \lambda_n(\theta)' g_i(\theta) \big|^3 \leq C \| \lambda_n(\theta) \|^3 = C \| \lambda^*(\theta) \|^3 + o_p(1), \end{aligned}$$

because $\sup_{y,\theta} \| g(y, \theta) \|^3 < \infty$; in view of (20) we know that $o_p(1)$ does not depend on θ. Now, continuity of $\| \lambda^*(\theta) \|^3$ at θ_0 proves the claim.

Step 2 (Approximating the logarithm): We claim that for any $\epsilon > 0$, there exists $\delta > 0$ such that $\| \theta - \theta_0 \| < \delta$ guarantees that

$$\big| \log\big(\frac{1}{n} \sum_{i=1}^{n} \exp(\lambda_n(\theta)' g_i(\theta)) \big) - \frac{1}{n} \sum_{i=1}^{n} \lambda_n(\theta)' g_i(\theta) - \frac{1}{2} \lambda_n(\theta)' \big(\frac{1}{n} \sum_{i=1}^{n} g_i(\theta) g_i(\theta)' \big) \lambda_n(\theta) \big| \leq \epsilon + o_p(1)$$

for $n \to \infty$, where $o_p(1)$ does not depend on θ.

First, recall that $\mid \log(x) - \log(y) \mid \leq C \mid x - y \mid$ when x and y are bounded away from zero. Therefore, for sufficiently large n, and θ close enough to θ_0, we have

$$\begin{aligned}
&\big| \log\big(\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda_n(\theta)' g_i(\theta))\big) - \log\big(1 + \frac{1}{n}\sum_{i=1}^{n} \lambda_n(\theta)' g_i(\theta) + \frac{1}{2}\lambda_n(\theta)'\big(\frac{1}{n}\sum_{i=1}^{n} g_i(\theta) g_i(\theta)'\big)\lambda_n(\theta)\big)\big| \\
&\qquad \leq C\big|\frac{1}{n}\sum_{i=1}^{n}\big(\exp(\lambda_n(\theta)' g_i(\theta)) - 1 - \lambda_n(\theta)' g_i(\theta) - \frac{1}{2}\lambda_n(\theta)' g_i(\theta) g_i(\theta)' \lambda_n(\theta)\big)\big| \\
&\qquad \leq C \parallel \lambda^*(\theta) \parallel^3 + o_p(1), \quad (23)
\end{aligned}$$

where the last inequality is due to step 1. Similarly, recall that $\mid \log(1 + x) - x \mid \leq C \mid x \mid^2$ when x is bounded away from -1. Hence, for sufficiently large n and θ close enough to θ_0, we also have

$$\begin{aligned}
&\big| \log\big(1 + \frac{1}{n}\sum_{i=1}^{n} \lambda_n(\theta)' g_i(\theta) + \frac{1}{2}\lambda_n(\theta)'\big(\frac{1}{n}\sum_{i=1}^{n} g_i(\theta) g_i(\theta)'\big)\lambda_n(\theta)\big) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n} \lambda_n(\theta)' g_i(\theta) - \frac{1}{2n}\sum_{i=1}^{n} \lambda_n(\theta)' g_i(\theta) g_i(\theta)' \lambda_n(\theta)\big| \\
&\qquad \leq C\Big(\frac{1}{n}\sum_{i=1}^{n} ||g_i(\theta)|| ||\lambda_n(\theta)|| + \frac{1}{2n}\sum_{i=1}^{n} ||g_i(\theta)||^2 ||\lambda_n(\theta)||^2\Big)^2 \\
&\qquad \leq C||\lambda_n(\theta)||^2 + C||\lambda_n(\theta)||^3 + C||\lambda_n(\theta)||^4 \leq C||\lambda^*(\theta)||^2 + o_p(1), \quad (24)
\end{aligned}$$

where we used $\sup_{y,\theta} ||g(y, \theta)|| < \infty$. Combining (23) and (24) with continuity of $\lambda^*(\theta)$ proves the claim.

Step 3 (Approximating the Betel density): Now, note that step 2 proves that the log of the Betel density is approximated by $-\frac{1}{2}\lambda_n(\theta)' Q_n(\theta)\lambda_n(\theta)$, where $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta) g_i(\theta)'$. That is, rewriting the result of step 2 yields

$$\mid \frac{1}{n}\sum_{i=1}^{n} \log w_i(\theta) + \frac{1}{2}\lambda_n(\theta)' Q_n(\theta)\lambda_n(\theta) \mid \leq \epsilon + o_p(1)$$

in some neighborhood of θ_0 as $n \to \infty$. Therefore, from (17), (20), and continuity of $Q(\theta) = E(g_i(\theta)g_i(\theta)')$, we can choose $\delta > 0$ such that $\parallel \theta - \theta_0 \parallel < \delta$ implies that

$$| \frac{1}{n}\sum_{i=1}^{n} \log w_i(\theta) + \frac{1}{2}\lambda^*(\theta)'Q(\theta_0)\lambda^*(\theta) | \leq \epsilon + o_p(1) \tag{25}$$

as $n \to \infty$; since $\sup_{y,\theta} ||g(y,\theta)|| < \infty$ and the convergence of $\lambda_n(\theta)$ is uniform, $o_p(1)$ does not depend on θ.

The remaining step is approximating $\lambda^*(\theta)$. Since $\lambda^*(\theta)$ is differentiable at θ_0, we know that there exists $\delta > 0$ such that $\parallel \theta - \theta_0 \parallel < \delta$ implies that

$$\parallel \lambda^*(\theta) - L(\theta_0)(\theta - \theta_0) \parallel \leq \epsilon, \tag{26}$$

where $L(\theta) = \frac{\partial \lambda^*(\theta)}{\partial \theta'}$. Since $\lambda^*(\theta)$ is implicitly defined from $E(g_i(\theta)\exp(\lambda' g_i(\theta))) = 0$, we use the implicit function theorem to obtain

$$L(\theta_0) = -Q(\theta_0)^{-1}\frac{\partial E\big(g_i(\theta)\exp(\lambda' g_i(\theta))\big)}{\partial \theta'}\Big|_{\theta=\theta_0,\lambda=0} = -Q(\theta_0)^{-1}\Gamma(\theta_0),$$

where $\Gamma(\theta_0) = \frac{\partial E(g_i(\theta))}{\partial \theta'}\big|_{\theta=\theta_0}$. Combining (25) and (26) proves the proposition.

REFERENCES

Chamberlain G, Imbens GW. 2003. Nonparametric applications of Bayesian Inference. *Journal of Business and Economic Statistics* **21**: 12–18.

Chernozhukov V, Hansen C. 2006. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics* **132**:491–525.

Chernozhukov V, Hong H. 2003. An MCMC approach to classical estimation. *Journal of Econometrics* **115**: 293–346.

Dunson D, Taylor J. 2005. Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics* **17**: 385–400.

Gelman A, Carlin JB, Stern HS, Rubin DB. 1995. *Bayesian Data Analysis.* Chapman and Hall: London.

Graddy K. 1995. Testing for imperfect competition in the Fulton fish market. *Rand Journal of Economics* **26**: 75–92.

Huber PJ. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**: 73–101.

Jaynes ET. 2003. *Probability Theory: the Logic of Science.* Cambridge University Press: Cambridge.

Jeffreys H. 1961. *Theory of Probability.* Clarendon Press: Oxford.

Kottas A, Gelfand AE. 2001. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association* **96**: 1458–1468.

Lavine M. 1995. On an approximate likelihood for quantiles. *Biometrika* **82**: 220–222.

Monahan JF, Boos DD. 1992. Proper likelihoods for Bayesian anaylsis. *Biometrika* **79**: 271–278.

Rubin D. 1981. The Bayesian bootstrap. *The Annals of Statistics* **9**: 130–134.

Schennach SM. 2005. Bayesian exponentially tilted mmpirical likelihood. *Biometrika* **92**: 31–46.

Yu K, Moyeed RA. 2001. Bayesian quantile regression. *Statistics and Probability Letters* **54**: 437–447.

van der Vaart AW. 1998. *Asymptotic Statistics.* Cambridge University Press: Cambridge.

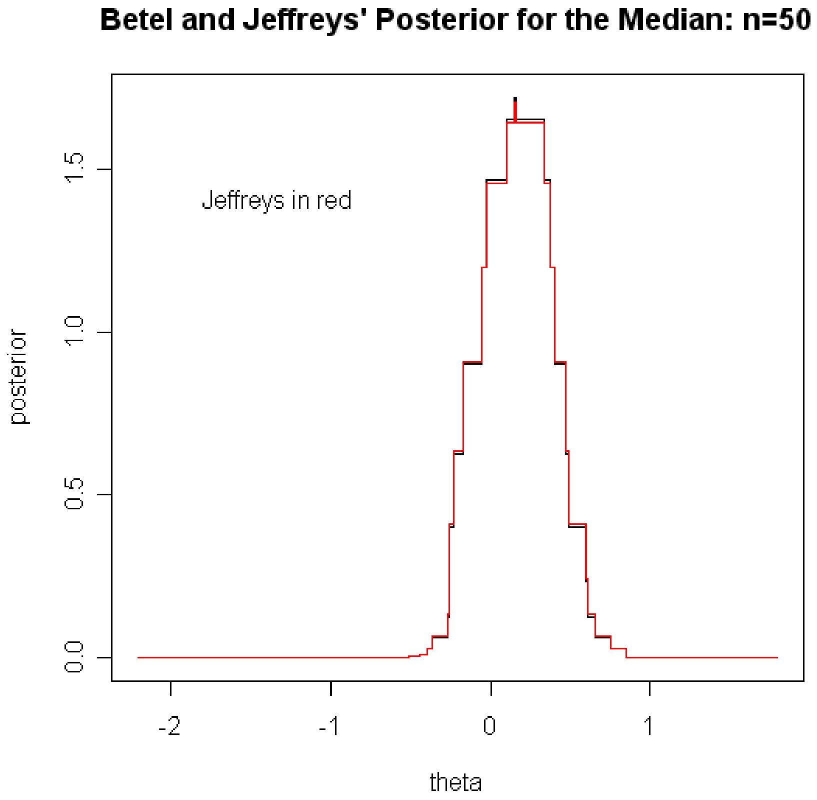


Figure 1: The figure shows Jeffreys' posterior for the median in red and the Betel posterior in black. The data are a random sample of size 50 from a normal distribution and in both cases a uniform prior was assumed.

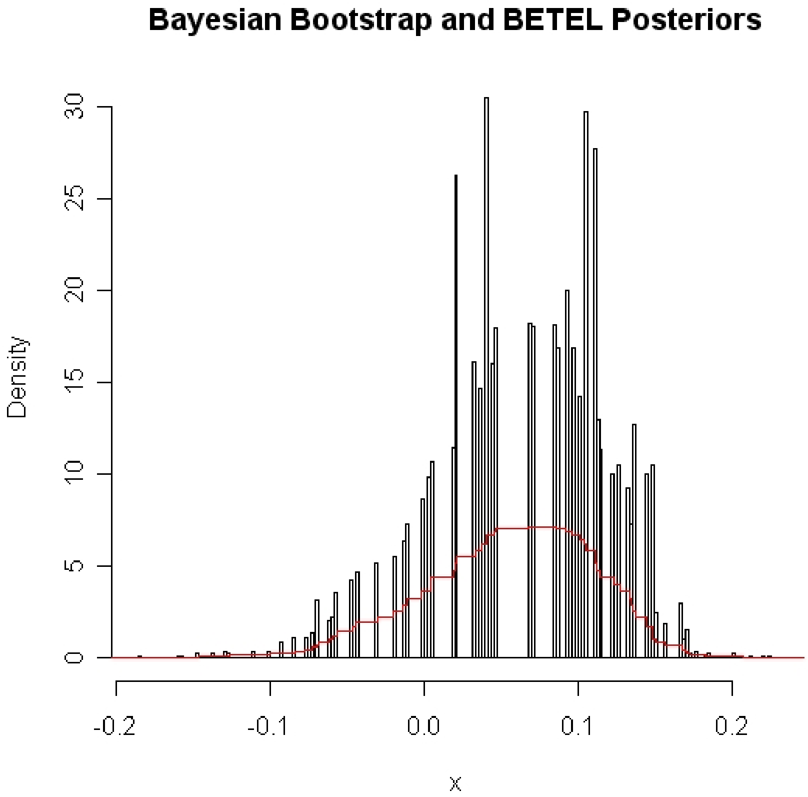


Figure 2: The vertical lines represent the discrete Bayesian Bootstrap posterior distribution while the red curve shows the Betel posterior. The data were 500 realizations from a standard normal distribution and the priors were uniform.

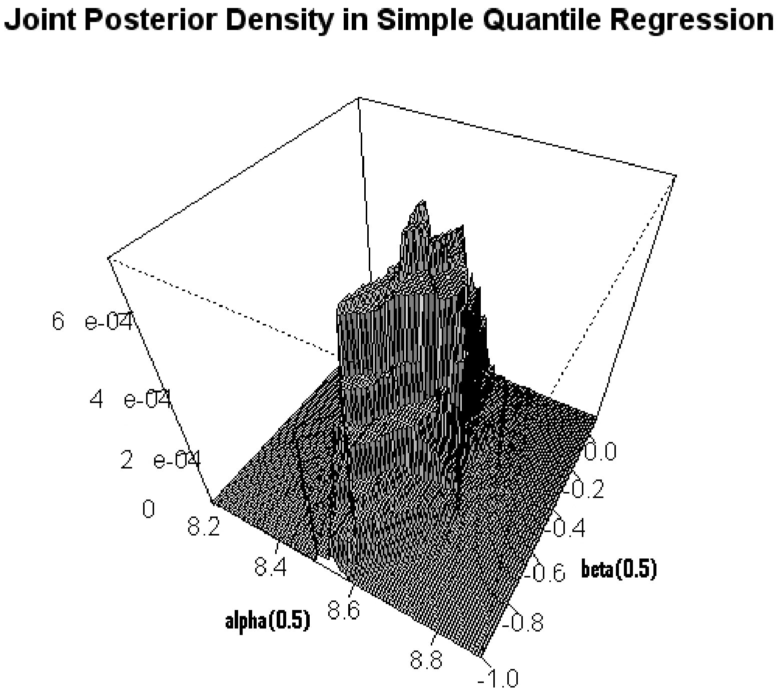


Figure 3: This is the joint posterior density of slope and intercept in the median quantile demand curve for fish. The data are Graddy's 111 observations on price and quantity traded. A uniform prior was assumed.

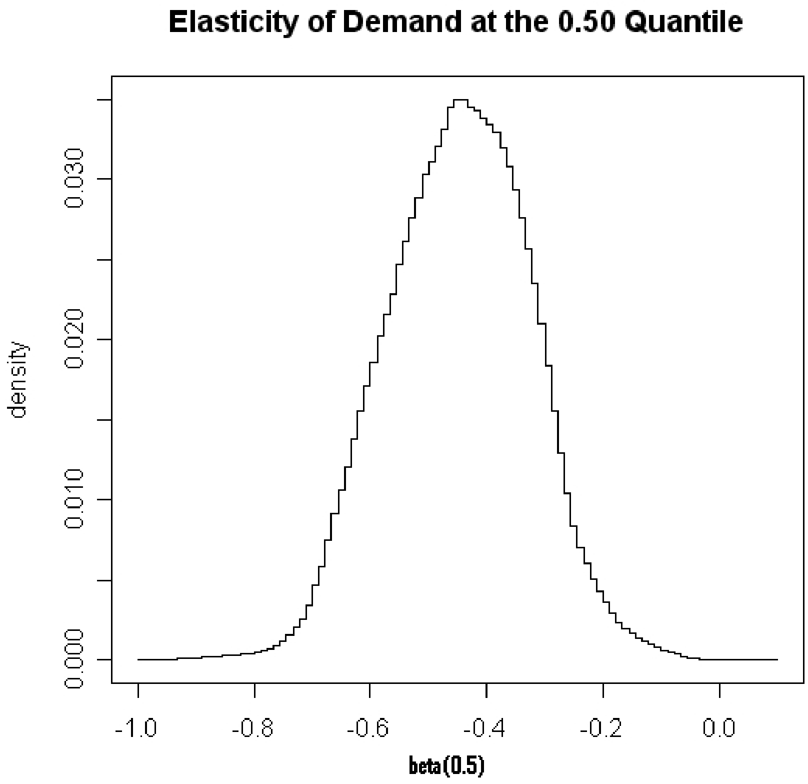


Figure 4: This is the marginal posterior density of the elasticity of demand for fish. It was calculated by summing the joint posterior density shown in figure 3 over the intercept rows.

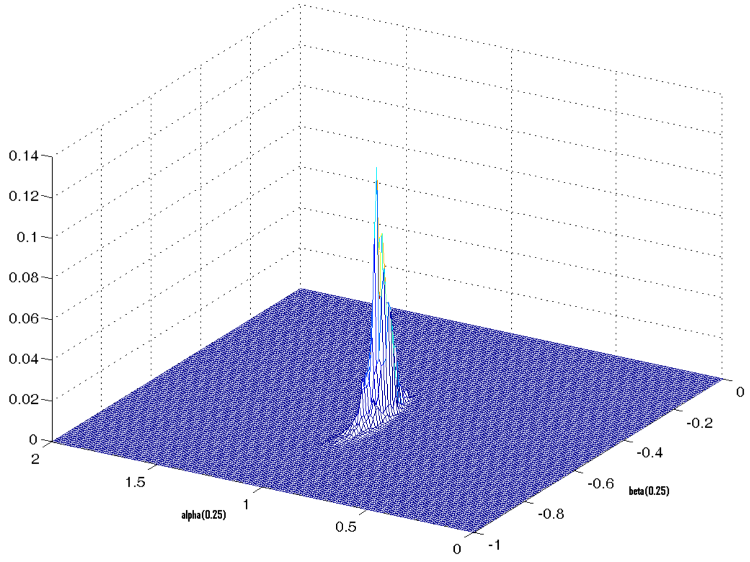


Figure 5: The figure shows the joint posterior density of slope and intercept in the 0.25 quantile regression using simulated data with $n = 500$ and eight strong instruments.

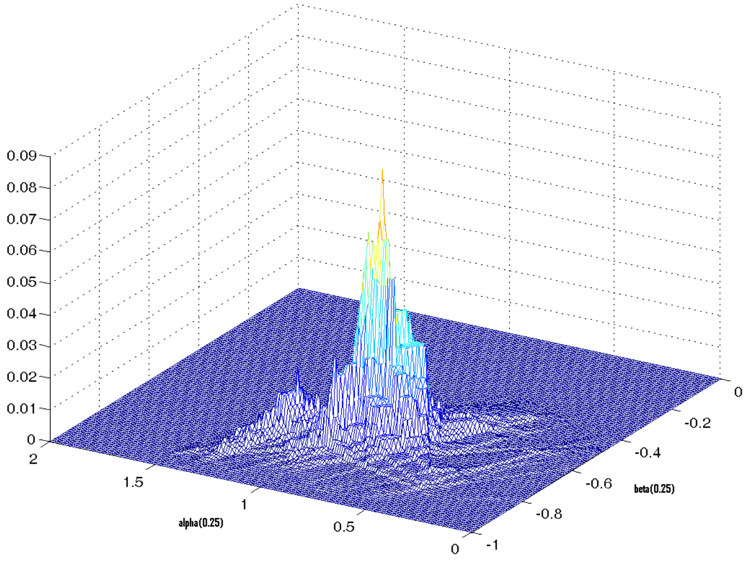


Figure 6: The figure shows the joint posterior density of slope and intercept in the 0.25 quantile regression using eight weak instruments. The data are simulated with sample size 500. Note the appearance of secondary modes.

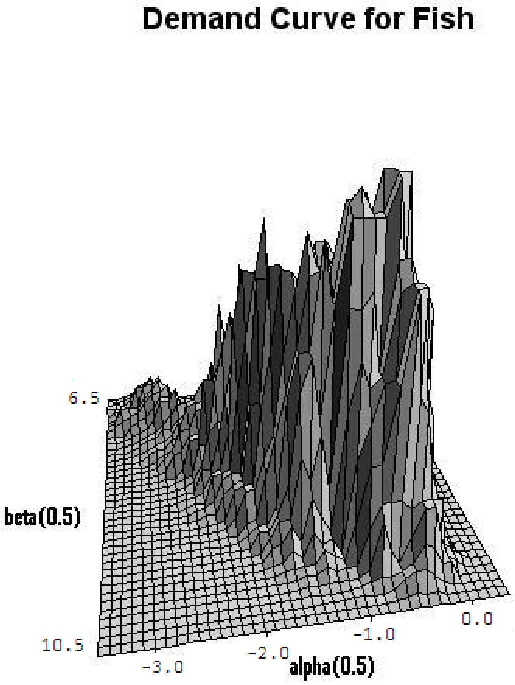


Figure 7: This is the joint posterior density of slope and intercept in the median demand curve for fish using two weather variables as instruments. The data are Graddy's with $n = 111$. Note the appearance of a potential minor mode; see also figure 8 that shows the marginal posterior.

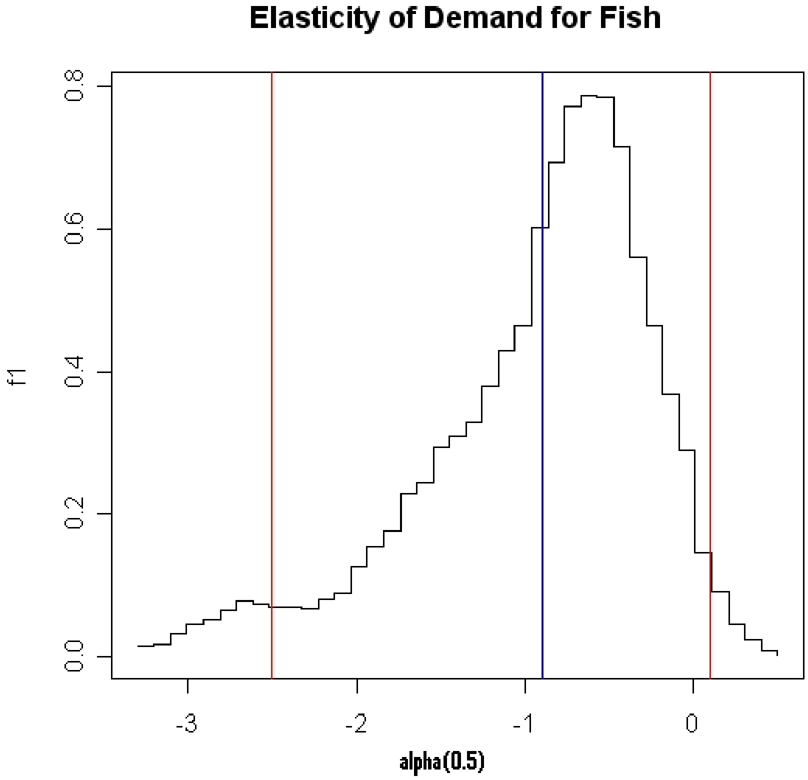
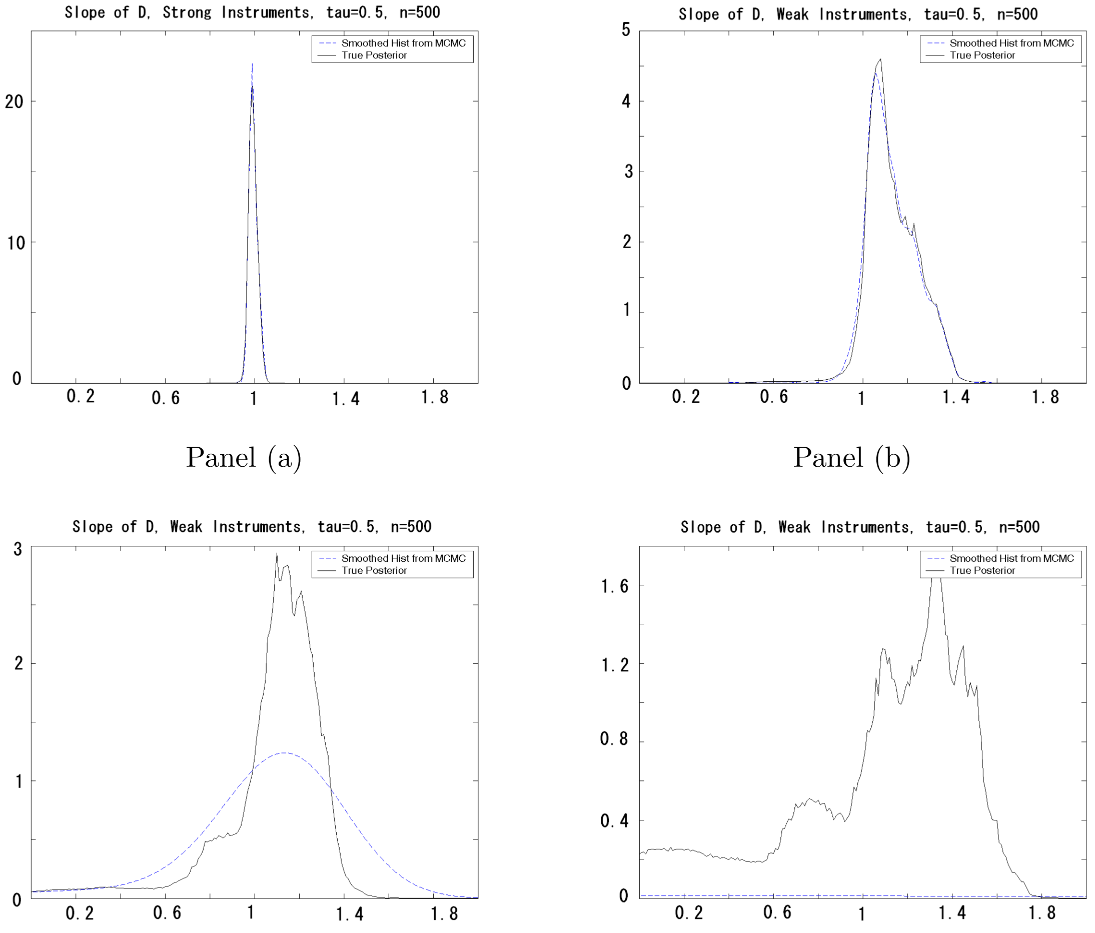


Figure 8: This is the marginal posterior density of the median elasticity of demand for fish found by summing out the intercept in the previous figure. The red lines mark a 95% highest posterior density interval and the blue line represents a previously reported frequentist point estimate.



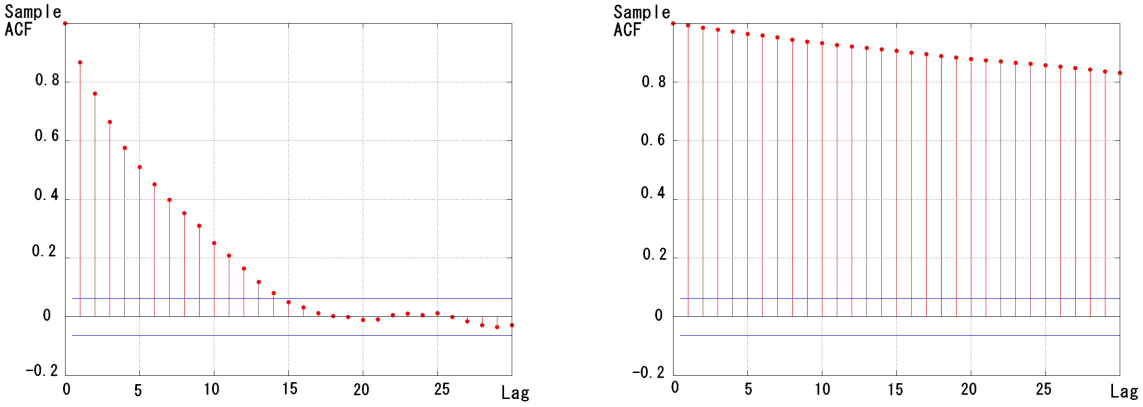
Panel (a)

Panel (b)

Panel (c)

Panel (d)

Figure 9: These figures show the true marginal posteriors and the MCMC samples under various qualities of identification. As instruments became weaker, the marginal posteriors showed multiple modes and heavy tails. Consequently, the MCMC samples could not get the posteriors with the given number of iterations.



Type I

Type II

Figure 10: These figures are two autocorrelation functions (ACFs) from the five chains run for the panel (c) in figure 9. Horizontal lines show 95% cofidence bounds for no autocorrelation. The fact that two completely different ACFs still appear after 10,000 iterations can be noted as an indication of the pathological shape of the target posterior.